Data Wrangling on Mutual Funds

Gourav K Verma

Data Science, Bellevue University

DSC540-T303 Data Preparation

Professor Catherine Williams

May 29, 2020

Abstract

A mutual fund is a company that pools money from many investors and invests the money in securities such as stocks, bonds, and short-term debt. The combined holdings of the mutual fund are known as its portfolio. Investors buy shares in mutual funds. Each share represents an investor's part ownership in the fund and the income it generates. There are many parameters and points attached to a Mutual Funds. Out of these I will be putting my focus on important factors tied to a high performing Mutual funds. For the final project data was gathered and cleaned from three data source- API, Web and Flat File(CSV).

Keywords: Data Wrangling, Data Science, Data Format, Data Source, API, Data Cleaning, Fund Portfolio, Morning star ratings.

Sources of Data

For the project data was gathered from 3 sources – API, Web and Flat file

Flat File: Mutual Funds.csv downloaded from
https://www.kaggle.com/stefanoleone992/mutual-funds-and-etfs

The file contains 25,265 Mutual Funds and 2,353 ETFs with general aspects (as Total Net Assets, Management Company and size), portfolio indicators (as cash, stocks, bonds, and sectors), returns (as year-*t*o-date, 2018-10) and financial ratios (as price/earnings, Treynor and Sharpe ratios, alpha, and beta).

API: https://financialmodelingprep.com/

Stock information and data

Website: https://finance.yahoo.com/mutualfunds?offset=0&count=100

Yahoo! Finance is a media property that is part of Yahoo!'s network. It provides financial news, data and commentary including stock quotes, press releases, financial reports, and original content. It also offers some online tools for personal finance management.

Relationship between sources will be stablished based on the **fund's code name**.

Flat File Cleaning Steps

- From CSV file data was read into ***Pandas*** Dataframe
- It had 125 columns and 25308 Rows, which was sorted and new dataframe was created with selected 45 important columns.
- Checked the % of NaN values in each column and for less than 10% Rows with NaN column values was removed. We left with 21750 rows after removing 14.06% of our dataframe.
- Columns were converted from object to numeric type.
- Some columns were renamed.
- Checked for outliers in column fund_return_10years
- Final Dataframe was written into a CSV file.

API Data Cleaning Steps

- A successful connection was made using ***urllib*** with API and all available Mutual Fund's information was retrieved.
- A blank Dictionary was used to store the JSON data retrieved from API. Later was converted to a readable pandas dataframe.
- Total 22 columns and 1505 rows were retrieved.
- In 11 columns had no data, which were removed from dataframe.
- Checked for duplicate entries using 'Symbol' column.
- Columns were formatted to correct data types.
- Final dataframe was loaded into CSV file.

Web Data Cleaning Steps

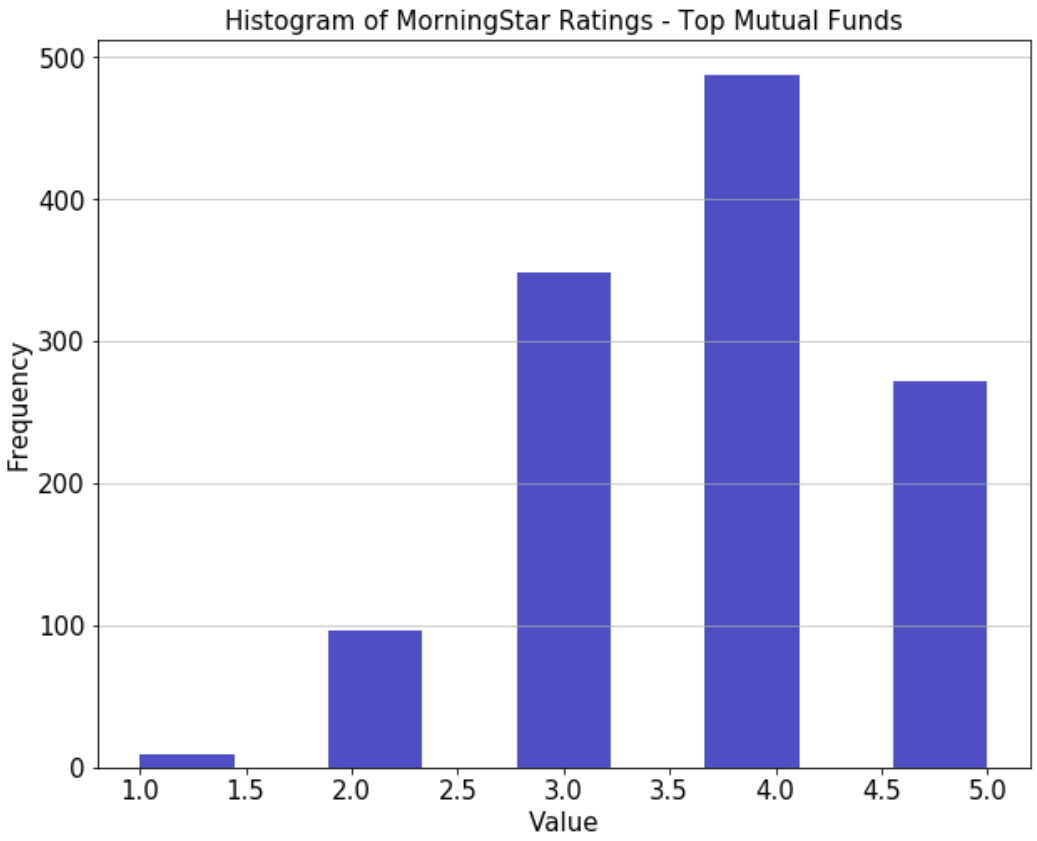
- Information about Top Mutual funds(Approx 1240) were read from yahooFinance.com using ***BeautifulSoup***
- Duplicate entries were removed using Symbol column.
- Columns were formatted to correct data types.
- Outliers were identified using box plot.
- Final DataFrame was loaded into CSV file.

Merging and Database load

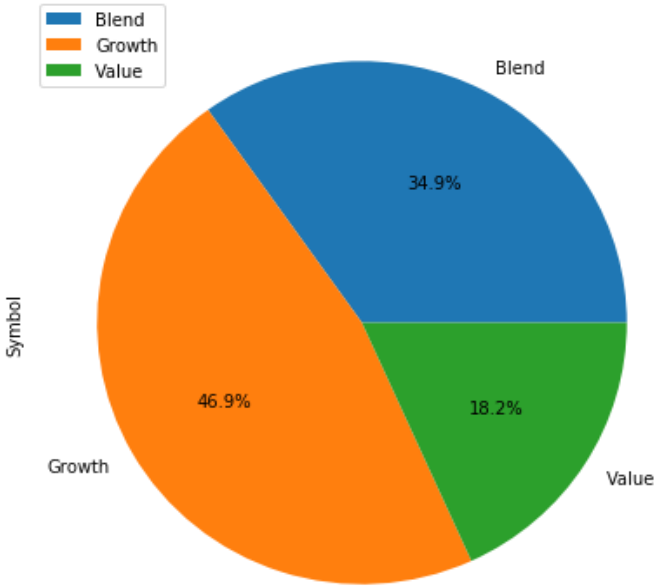
- The 3 CSV from different sources were merged with ‘inner’ join, for having subset of high performing Mutual Funds.
- I used DB Browser (SQLite) on my system and created database DSC540_Prj.db
- Inside this database a table 'Mutual_Fund' has been created by loading the final CSV file 'MF_all.CSV'
- Then data was read from database to create Visualizations.

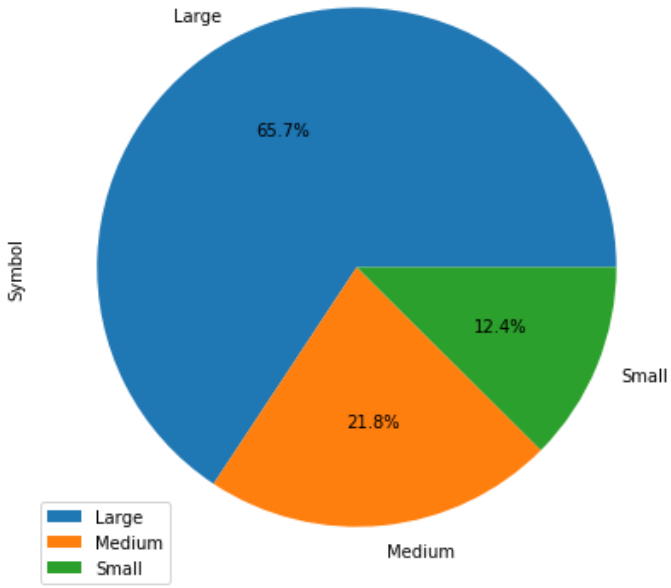
Visualization

- I am interested to know what are the MorningStar ratings of these top Mutual Funds. Hence, plotted histogram of it.
- Approx. 50% of Mutual Funds has MorningStar ratings of ‘4’.

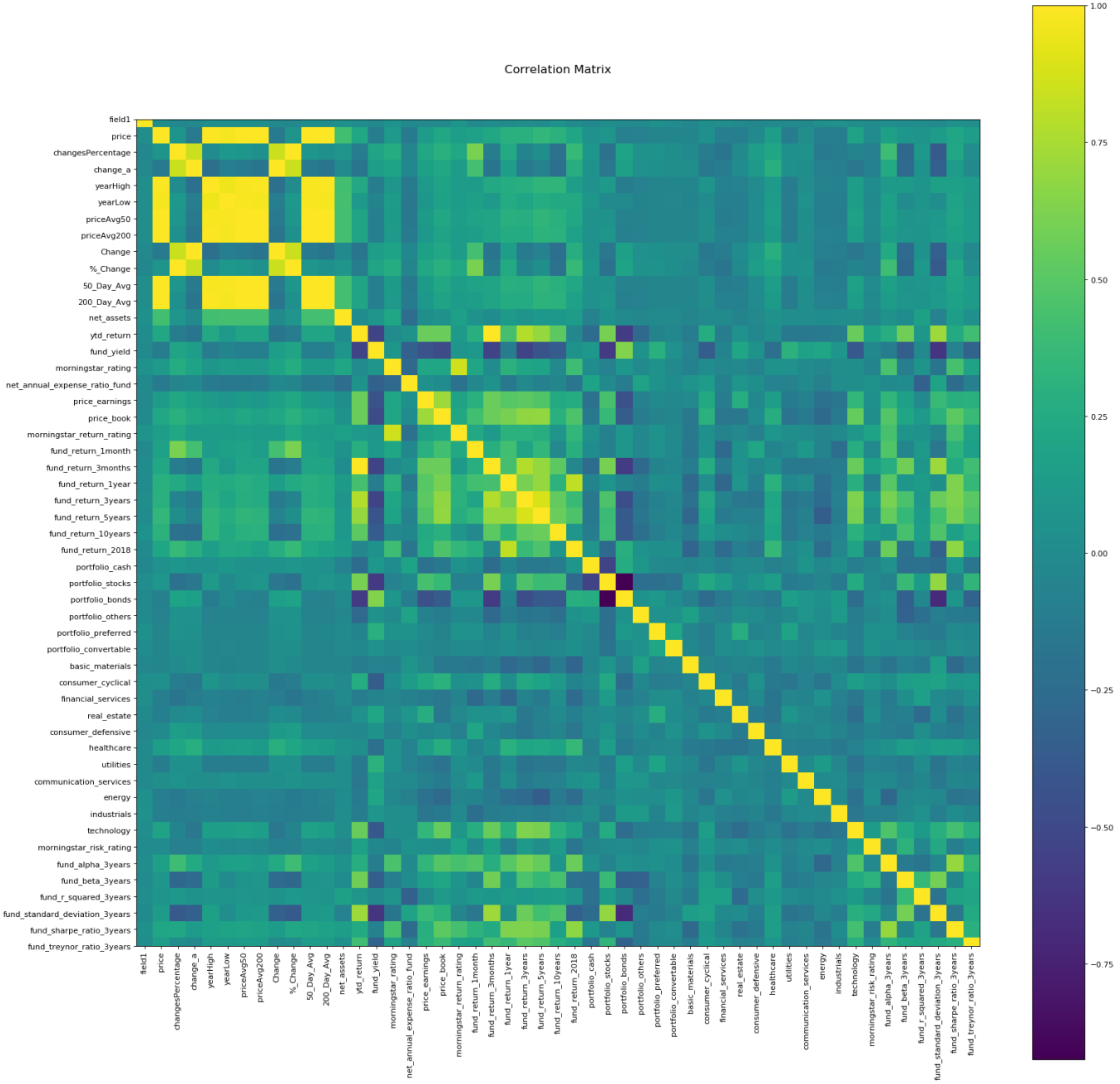


- We would also like to know % of the portfolio types and fund types distribution. Below pie charts strikes the difference.
- We can say **Large-Growth Fund** types are high performing.

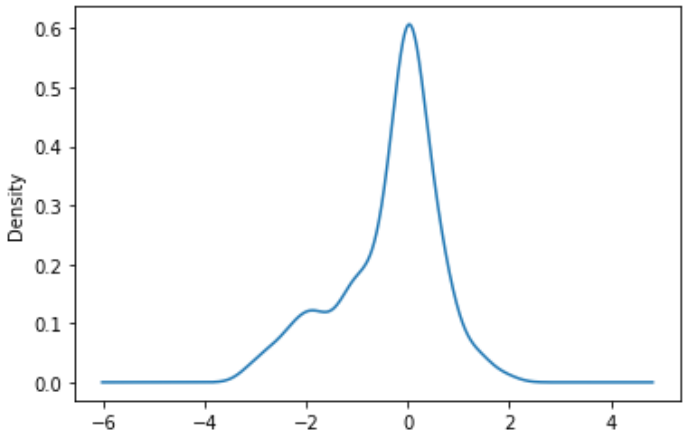




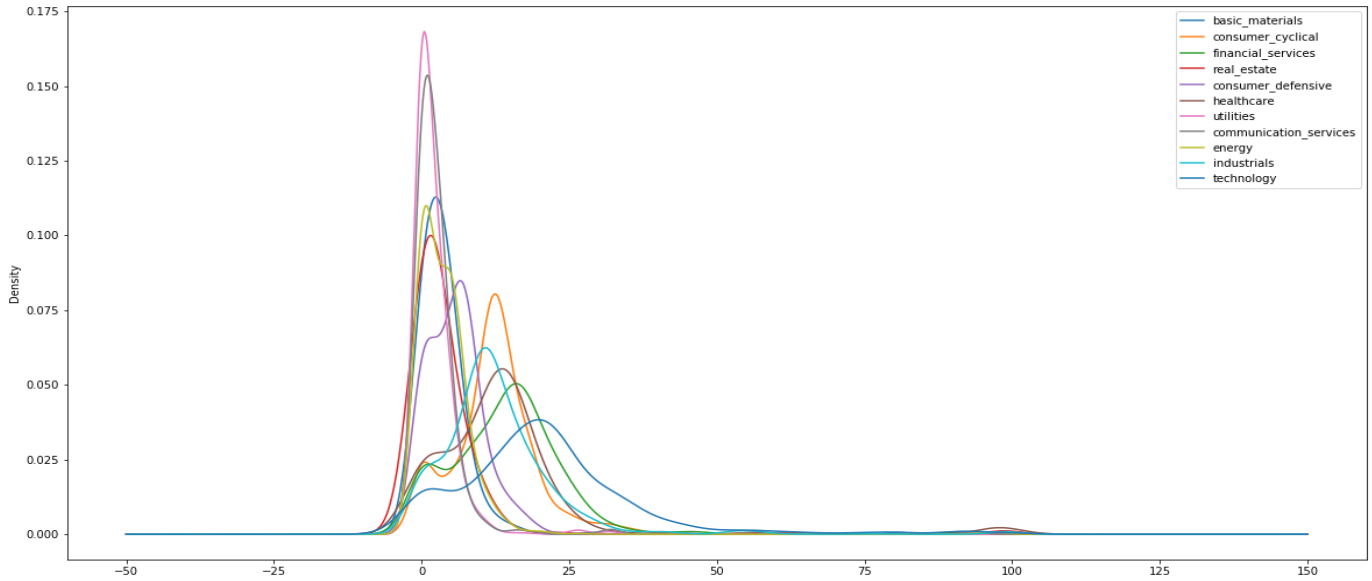
- Below plot shows the correlation Matrix of all the variables in the Dataframe. Lighter spots has higher correlation.



- Below density plot shows the % Change in a normal day. On an average most of the funds have daily changes near to 0.00 or Less.



- Below plot shows the density of different sectors of Mutual Fund's portfolio distribution.
- Most of the funds has investment in Technology, Financial Services, Healthcare and Industrials.



References

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*.

Sarkar, T., & Roychowdhury, S. (2019). *Data Wrangling With Python: Creating actionable data from raw sources*. Birmingham, UK: Packt Publishing.